DOCUMENT RESUME

ED 219 450 TM 820 478

AUTHOR Scheuneman, Janice Dowd
TITLE Item Bias and Test Scores.
PUB DATE Mar 82
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Latent Trait Theory; Research Methodology; *Research Problems; Simulation; *Test Bias; *Test Items

ABSTRACT

The connection between item bias and test scores was investigated using a simulation approach. Two samples of hypothetical examinees were simulated using an item response theory model. The two samples were identical, except that the mean theta value 1 sample was 5 less than the other. The simulated tests consisted of 50 items with characteristics which varied with the difficulty of the test for the group, the distribution of 6 values, the mean discrimination of items, and the condition of bias. The study resulted in several hypotheses concerning desirable characteristics of tests given that the actual amount and condition of bias in the tests is unknown: (1) the differences in true score between groups was larger with the peaked distribution than with the uniform distribution; (2) tests with relatively low discrimination appear to produce larger differences; (3) tests which are relatively easy for the high scoring group should have more rectangular distributions; (4) tests which are quite difficult seem to produce smaller differences when they are peaked; and (5) for tests with medium or high average discrimination, tests with difficulty centered between the groups appear to be least likely to produce inflated differences due to bias. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219450

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Item Bias and Test Scores

Janice Dowd Scheuneman
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Scheuneman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 478

A paper in J.D. Scheuneman (chair), Whither item bias research? A symposium presented at the annual meeting of the National Council on Measurement in Education, New York, March 1982.

Differences in test performance between groups defined by sex, ethnic background, age, SES or other such characteristics, and the issues arising from these differences have been of interest to psychologists and educators almost from the inception of objective mental testing. It was not, however, until the late 1960's and early 1970's that the issue of test bias rose almost to the status of a cause celebre. In the spirit of that time, test publishers and others concerned with the design and construction of tests became interested in detecting biased items during the test development process. The research then, however, tended to focus on the use of intact tests with reference to various criteria of performance such as college grades or supervisory performance ratings. For tests under construction, these methods came too late in the development cycle to permit changes or modifications in the test. Further, appropriate criterion measures were often unavailable or too expensive to collect. In consequence, techniques were devised to evaluate tests at the item level which did not require outside criterion measures. These were the procedures which have come to be known as item bias methods.

As time passed the area of test bias research became a hotbed of controversy. Darlington and Thorndike both pointed out serious methodological deficiencies with the commonly accepted regression definition of bias. More sophisticated statistical models were developed and proposed by Nancy Cole, Mel Novick and others. The debate was further heightened by the growing interest of the legal community who sought a model to receive their stamp of approval for discrimination cases. This controversy over definitions of fairness and methods of establishing the existence or non-existence of bias eventually

culminated in a special issue of the Journal of Educational Measurement in 1976. Oddly, this journal issue seemed almost to terminate the discussion with little new material appearing in the literature since then.

The infant field of item bias meanwhile continued to grow, despite criticisms that these methods were essentially circular. Because no external criterion measures were used, the items were of necessity judged to be biased in relation to other items on the test or in relation to the test score or a true score estimated from item performance. Nonetheless, these methods have stood up well to further investigation. While the test bias methods were foundering in a morass of definitional and methodological controversy, item bias studies were quietly churning out information which I believe will eventually lead us to some of the answers we are seeking about what bias is and how it operates. These methods are basically circular, however, defining ability as what the test measures, and as such are fundamentally incapable of addressing the key question: How much, if any, of the observed performance differences between groups can be accounted for by bias in the test. (See Cole, 1981, for a brief overview and discussion of bias research.)

I would like to suggest as a further direction for research that we begin to become concerned again about making a connection between bias in test items and bias in test scores, permitting us eventually to reapproach the issue of fairness or bias in test use from a new perspective. In this paper, I will take a step in that direction by investigating the connection between test scores and item bias using a simulation approach. With these techniques, the kind and amount of bias in the items and characteristics of the test such as its difficulty for the groups of interest can be systematically varied and the effect on test score differences observed. Specifically, the

purpose of the study is twofold. First, bias effects, when isolated, have often been dismissed as inadequate to produce the observed performance differences. Hence, we will be interested in determining how much difference can be produced as the result of biased items and whether different bias effects are additive or interact in ways which increase the observed differences beyond what either would produce alone. Second the results under different item and test characteristics will be summarized in order to determine if guidelines can be suggested for the design of tests which might minimize the inflation of score differences due to bias when we do not know the true state of bias in the test. I will first introduce a model or framework in terms of which I will define my terms and present my rationale for the simulation design. I will then present the results of this study and the implications of these results for our work in item bias. I will then describe my own plans for completing this study and suggest other research questions which seem to follow from this line of thought.

Theoretical Framework

The term "bias" is used statistically to indicate that the expected error of measurement is not zero. Used in this sense, to say a test is biased against a group is to say that the test systematically underestimates the ability of members of that group. The basic premise in this study is that tests are basically valid for members of different minority groups in that the tests measure the same underlying constructs which (given sufficient command of the English language) are then related to performance measures of interest in the much the same way for minorities as for Whites. The tests may still be biased in this statistical sense, however, in that they may be

yielding scores which are underestimates of the minority examinee's true ability. The reasons for such underestimates would lie in relatively stable characteristics of a test, such as the format in which the items are presented, or in idiosyncratic characteristics such as wording or content which will differ from item to item and, most likely, from one form to the next of a particular test.

In order to illustrate this conception of bias more clearly, let me suggest a true score model as follows:

$$X = \theta + \beta + \delta$$

where X = observed score
θ = true score
β = a bias component
δ = normal measurement error

In theory, the measurement error δ has a mean of zero, so if there were no bias component, the observed score mean would equal the true score mean. In statistical terms, this means that the observed scores are unbiased estimates of true scores. Suppose now that β, the bias component has a non-zero mean. In this case the observed score mean would not equal the true score mean. If, for a specific group, the mean of β is negative, the observed scores on the average will be under-estimates of the true scores.

Let now assume that β is different for different people and different items and has a mean of each item across the people in a given group and a mean and standard deviation of these item means across the items of a test. The magnitude of β will be influenced both by the stable characteristics and the idiosyncratic characteristics of the items in a test with the stable characteristics contributing primarily to the mean of β and the idiosyncratic to the variation. (See Scheuneman, in press, for a fuller development of this model.)

Further, this bias component is postulated to affect performance on specific test items in ways which will distort the probability of a correct response for persons of a given ability in ways associated with group membership. Such distortions can be easily represented in terms of item response theory (IRT) or latent trait theory where an unbiased item can be defined as one for which the item characteristic curve is the same for members of different population subgroups. The item characteristic curve is a mathematical expression relating the probability of a correct response to an item with the ability of the examinee. Most often in the measurement literature this relationship takes the form of a logistic function with three parameters, the graph of which closely resembles a normal ogive. These parameters specify the inflection point of the curve, b, which represents the difficulty of the item; the slope of the curve at the inflection point, a, which represents the discrimination of the item; and the lower asymptote of the curve, c, which is often thought of as the probability of getting the item correct by guessing. (See Lord, 1980, for details of IRT models and applications.) For simulation purposes it is possible to create a biased item by altering one or more of the parameters of the item characteristic curve for members of one group, hence distorting the probability of a correct response at given levels of ability.

Method

The simulation of test scores using an IRT model requires specification of two samples of hypothetical examinees in the form of two distributions of true ability values, θ, and the specification of item parameters for a set of hypothetical items which would make up the test. Given these four values, the probability of a correct response to any one of the items for any one of the examinees can be determined by finding the value of the item characteristic

function with the specified parameters at the specified value of θ. When these probabilities are summed across items for a given examinee, the result is that examinee's true score on that test or item set. Summing the probabilities across people for a given item provides the true difficulty of the item for that sample of examinees. Both values were obtained in this study, but the discussion in this paper will focus on results in terms of true scores. Notice that these results consist only of successive solutions for the value of the item characteristic curves at specified values; no random response variation is included in these analyses.

A distribution of theta values was randomly generated to have a mean of .5 and a standard deviation of 1 on an arbitrary IRT scale. This distribution represented the abilities of the majority group. The minority group was assumed to be less able on the average due to differences in opportunity to learn, value placed on academic learning in the culture and other factors in the environment or culture which affect test scores in ways which would be generally recognized as valid for the purposes of the test. In order that results not be influenced by any distributional characteristics other than the mean, this second group was generated by subtracting one from each of the theta values in the first sample. Hence, this sample is identical to the first in all distributional characteristics, except the mean which was then -.5. I will treat these groups as prototypes of White and Black examinee groups although the second group could equally well represent another ethnic minority group or the two groups could be differentiated by sex, SES, age or other group characteristics where large score differences are observed on some particular test. The assumption here is that some true difference exists for the reasons given but that this diffrence can be magnified or

inflated by the action of bias. Remember also, that there are no real items. The bias component of the observed scores, which is associated with group membership, is assumed to be reflected by the distorted probabilities resulting from the changes in the item parameters in the simulation design.

The tests each consisted of fifty items with characteristics which were systematically varied as part of the simulation experiment. The characteristics were selected from among those in earlier studies using IRT models to simulate bias (Pine & Weiss, 1976, and Rudner, Getson, & Knight, 1980) and from those shown in preliminary analyses to be most likely to have an effect on the test scores. The characteristics chosen for study were:

The difficulty of the test for the group. Three difficulties were examined, where tests had mean difficulties (mean b's) of .5, 0, -.5. That is, the three tests were centered in difficulty around the mean for the majority group, between the two group means, and centered around the mean for the minority group.

The distribution of b values. In their study Pine and Weiss (1976), suggested that the kurtosis or flatness versus peakedness of the distribution of b's affected the apparent bias in the scores. All analyses were, therefore, produced for two distributions of item difficulties. One distribution of b's was generated from a normal distribution with a mean of 0 and a standard deviation of 1. The second distribution was generated from a uniform distribution of random numbers, with a range of -3 to +3, adjusted to the same mean as the peaked distribution. Note, however, that the uniform distribution is also more variable than the peaked and similar results may have been obtained from peaked distribuitons with much greater variability. Further analyses would be required to determine which of these factors, variance or kurtosis, is the more important to the results.

The mean discrimination of items. Preliminary analysis showed scores to be affected by both the mean and standard deviation of the a values. Consequently, results were evaluated for three different distributions of a's. All were normally distributed with means of .30,.70, and 1.0 with standard deviations equal to one half the value of the mean.

Condition of bias. Bias was generated by adding a random factor to one of the parameters. The conditions examined were 1) bias added to the b parameters for Blacks with the mean bias causing the items to be more difficult, 2) bias added to the c parameters for Whites with the mean bias making the items easier to guess at correctly, and 3) bias added to the a parameter for either Whites or Blacks with the mean of the biased a's the same as for the unbiased or greater for one of the two groups. Bias could be attached to one, two, or all three of the parameteres for a particular hypothetical test.

Results

Generation of distributions

The distributions of parameters were generated by selecting random values from a normal distribution with the specified mean and standard deviation. While the desired mean values could be obtained for any sample drawn by subtracting a constant from all numbers in that sample, the standard deviation was more difficult to match with only 50 parameter values. Consequently successive samples were drawn and the standard deviation computed until a sample was selected where the desired value was approximated. For the uniform b distribution, samples were successively drawn from a rectangular distribution with a range of -3 and +3 in order to obtain a low kurtosis value. The obtained means and standard deviations for the different parameter distributions and the theta distribution are shown in Table 1.

Table 1

Summary of Generated Values[1]

	Mean	S.D.	Kurtosis
Abilities:			
θ - Majority	+.50	1.00	
θ - Minority	-.50	1.00	
Parameters:			
Unbiased:			
a - low	.30	.15	
a - medium	.71	.32	
a - high	1.01	.48	
b - peaked	.002	.96	2.42
b - uniform	.002	1.82	1.59
c	.20	.048	
Bias Quantities:			
a - low	.003	.14	
a - medium	.003	.32	
a - high	.003	.48	
b	.50	.50	
c	.025	.049	
Biased Parameters:			
a + a - low	.30	.20	
a + a - medium	.71	.44	
a + a - high	1.01	.74	
b + b - peaked	.50	1.10	3.03
b + b - uniform	.50	1.89	1.82
c + c	.225	.071	

[1]There were 500 ability values generated for each group and 50 values for each of the parameter sets.

[2]The mean of the b's varied according to condition. Values used were -.50, 0, +.50.

[3]The mean of the a bias quantities varied according to condition. These were zero or plus or minus one s.d. of the respective unbiased a.

Bias was introduced into the items by adding a random term or bias quantity to each of the generated item parameter values. Therefore, when the probability of a correct response was computed for the item characteristic curve at a given ability value, the new, "biased" parameters were used for the group to which they were attached instead of the original unbiased parameter values. The bias quanities were also generated by drawing randomly from a normal distribution with mean zero and a standard deviation equal to that of the original parameters for the a's and c, and equal to .5 for the b's. (The same set of bias quantities was used for both the peaked and uniform distribution of b's). These bias quantities will be denoted by Δ a, Δ b, and Δ c. The new biased parameters, the original parameter plus the bias quantity, had the same mean as the original parameters but a larger standard deviation. The means and standard deviations of the bias quantities and of the biased parameters are also shown in Table 1.

The first of the preliminary analyses explored the effect on true score of the biased parameters when the mean of the bias quantity for a given parameter was zero. The zero mean essentially implies that bias in one direction is cancelled out by bias in the other direction. Indeed, in the case of the b and c parameters, the true scores resulting from the biased parameters differed only slightly from those resulting from the unbiased parameters in either mean or standard deviation. Consequently, the condition with the mean bias quantity of zero was not explored further for either the b or c parameters. For the a parameters, however, the true scores resulting from the biased a terms are different from those resulting from the unbiased a terms, even when the mean of Δ a was zero. This condition is explored, therefore, both when the Δ a is attached to the a's for the Whites and where it is attached to the a's for the Blacks.

Of particular interest in this study, however, is the case where the bias effects do not cancel one another out, that is, when the mean bias quantity is not zero. Therefore, a constant term was added to bias quantities in order to provide a mean greater than zero. For the $\underline{g}$ and $\underline{c}$ parameters, means less than zero were also obtained in the same manner. For the bias in $\underline{b}$, a constant of .5 was added to all Δ b values. This bias was considered only as attached to Blacks. (The same quantity subtracted from the mean and attached to the White group parameters was considered but the results were essentially identical.)

For the $\underline{c}$ parameters, the mean was moved plus or minus one-half the standard deviation of the unbiased $\underline{c}$'s. Again the sign of the mean bias quantity or the group to which the bias terms were attached resulted in essentially the same difference in true scores. The single condition, the positive mean attached to the majority group, was chosen because higher $\underline{c}$'s for Whites seemed more in line with expectation and because subtracting the Δ c's from the Black parameters resulted in some negative biased $\underline{c}$'s. The magnitude of the difference, however, was the same for all four conditions and could be computed on the basis of the information generated if desired.

The relationships of the $\underline{a}$ parameters turned out to be more complex so that the full range of possibilities was explored. True scores differences were obtained when bias was attached to either group and when means of the bias terms were positive, negative, or zero, for each of the three groups of unbiased $\underline{a}$ parameters. The bias quantities were separately generated for each set of $\underline{a}$'s and the means were equal to zero or to plus or minus one half the standard deviation of the respective unbiased parameter. Where Δ a's were subtracted from the original parameters, the sums were checked so that

when negative sums resulted, the Δa terms were modified so that the sum was zero.

Computation of true score differences

The differences in true score between the Black and White groups were first computed for the unbiased conditions. In general, the differences in true score units tended to be larger as item discrimination increased but the standard deviation of these scores also increased. In standard deviation units, therefore, the differences were largest with low discriminating items and decreased as the discrimination increased. The score differences were also larger when the test was relatively easy for the majority group and centered in difficulty for the minority group and with peaked distributions of b parameters. With uniform distributions, the differences were all very slightly less than the true difference of one standard deviation, ranging from .90 to .98 standard deviation units. With peaked distributions the score differences varied from .91 to 1.06 standard deviation units, but here the differences interacted with overall difficulty for the group and discrimination of the items. For the relatively easy test (mean b = -.5), the differences were larger with more discriminating items, while with the relatively difficult test (mean b = +.5), the differences were larger with the less discriminating items. Size of the true score differences in s. d. units is shown on Table 2.

Next, the true scores were obtained when bias was attached to only a single parameter at a time. The bias in the b parameter resulted in added differences in s.d. units close to the mean of .5 of the bias term. These ranged from .36 to .52 additional standard deviation units of true score

Table 2

Effects of Bias in One Parameter on
Score Differences Between Blacks and Whites

Peaked Distribution

Mean Discrimination	Mean b	Unbiased Condition[1]	Bias Effects[2]							
			b only	c only	a only for Whites			a only for Blacks		
					same	a > W	a > B	same	a > W	a > B
Low	+.50	.02	.48	.15	.01	.03		-.01		-.02
	0	-.01	.44	.15	.03	-.06		-.03		.06
	-.50	-.05	.41	.18	.05	-.11		-.05		.13
Medium	+.50	.06	.52	.09	-.02	-.02	-.02	.03	.02	.03
	0	.00	.43	.10	.00	-.07	.09	.00	-.07	.07
	-.50	-.06	.37	.12	.02	-.11	.18	-.03	-.15	.10
High	+.50	.06	.51	.09	-.08		-.08	.05	.04	
	0	-.02	.42	.09	.01		.12	-.03	-.10	
	-.50	-.09	.36	.10	.10		.32	-.10	-.23	

Uniform Distribution

Low	+.50	-.02	.48	.16	.02	.03		-.01		-.02
	0	-.02	.45	.17	.05	-.03		-.06		.01
	-.50	-.03	.44	.18	.08	-.06		-.09		.06
Medium	+.50	-.02	.41	.08	-.01	.01	-.04	.01	.02	-.02
	0	-.06	.40	.10	.04	.01	.12	-.03	-.07	-.01
	-.50	-.08	.40	.11	.09	.01	.24	-.06	-.18	.00
High	+.50	-.06	.37	.08	-.08		-.23	.04	.13	
	0	-.09	.35	.10	-.01		-.03	-.02	-.02	
	-.50	-.10	.36	.12	.10		.23	-.10	-.19	

1. Difference in S.D. units from the 1 unit of true difference. Negatives indicate differences less than one.

2. Bias effect is the added difference in S.D. units beyond that found in the unbiased condition.

difference for peaked distributions and from .35 to .48 for uniform distributions. For uniform distributions this difference was largest for the lower discrimination items, but showed little difference as a result of mean difficulty of the test. With peaked distributions, the relative difficulty of the test was more important in score change than the overall discrimination of the items, with the relatively easy test showing a larger difference than the relatively difficult. Table 2 shows the increase in the difference in true score beyond that shown in the unbiased condition.

The effect of larger c values for the majority group was much less marked than that of the b differences. There was very little difference here between the peaked and uniform distribution with additional true score differences ranging from .09 to .18 s.d. units for peaked and .08 to .18 for uniform. Differences tended to be larger for more difficult tests and less discriminating items in both cases.

In the case where bias was only in the a parameters, the differences tended to be small. When the mean bias quantity was zero, differences for both peaked and uniform tests ranged from -.08 to .10 when attached to Whites and tended to increase with the difficulty of the tests. When the bias was attached to Blacks the differences ranged from -.10 to .05 and decreased as the tests became more difficult. When the mean bias was not zero, the pattern depended less on the group to whom the bias was attached and more on the group for whom the items became more discriminating. When the mean discrimination was higher for Whites the differences became smaller as the test became more difficult, ranging from -.11 to .03 when attached to Whites and from -.23 to .13 when attached to Blacks. When the mean discrimination was higher for Blacks, the differences became larger as the test became more

Table 3

Effects of Bias in Two Parameters on
Score Differences Between Blacks and Whites

Peaked Distribution

Mean Discrimination	Mean b	b & c[1]	a & b[2] a on Whites same	a on Whites a > W	a on Whites a > B	a on Blacks same	a on Blacks a > W	a on Blacks a > B	a & c[3] a on Whites same	a on Whites a > W	a on Whites a > B	a on Blacks same	a on Blacks a > W	a on Blacks a > B
Low	+.50	.66	.52	.45		.62		.68	.16	.14		.14		.13
	0	.63	.49	.33		.54		.68	.18	.08		.12		.21
	-.50	.60	.48	.24		.47		.68	.23	.03		.13		.31
Medium	+.50	.66	.53	.46	.62	.52	.44	.60	.08	.06	.11	.12	.11	.12
	0	.59	.46	.33	.63	.41	.29	.54	.11	.02	.27	.10	.03	.18
	-.50	.50	.41	.23	.64	.33	.17	.50	.15	.00	.40	.09	-.03	.22
High	+.50	.64	.49		.60	.49	.40		.01		.02	.13	.03	
	0	.55	.48		.69	.32	.19		.11		.28	.06	-.01	
	-.50	.45	.49		.79	.19	.02		.21		.53	.00	-.14	

Uniform Distribution

Mean Discrimination	Mean b	b & c[1]	a & b[2] a on Whites same	a on Whites a > W	a on Whites a > B	a on Blacks same	a on Blacks a > W	a on Blacks a > B	a & c[3] a on Whites same	a on Whites a > W	a on Whites a > B	a on Blacks same	a on Blacks a > W	a on Blacks a > B
Low	+.50	.66	.53	.46		.57		.60	.17	.04		.15		.14
	0	.65	.54	.39		.49		.61	.21	-.13		.11		.18
	-.50	.64	.55	.34		.43		.63	.27	-.17		.09		.24
Medium	+.50	.56	.44	.41	.51	.41	.35	.43	.08	.09	.09	.09	.11	.06
	0	.54	.48	.41	.64	.37	.18	.44	.14	.11	.24	.07	.02	.09
	-.50	.52	.53	.40	.78	.34	.16	.45	.21	.12	.40	.05	-.07	.11
High	+.50	.52	.32		.27	.37	.39		.02		-.09	.15	.22	
	0	.51	.39		.47	.26	.19		.11		.12	.08	.08	
	-.50	.47	.51		.75	.17	.01		.23		.40	.01	-.08	

[1] Bias quantity is attached to b for Blacks and to c for Whites.

[2] Bias quantity in first case is attached to a for Whites & b for Blacks; in second case on both

difficult with values ranging from -.23 to .32 when attached to Whites and -.02 to .13 when attached to Blacks. In all cases, the rate of change from easy to difficult tests was largest with highly discriminating items, so that both the largest and smallest differences occurred on those tests.

If both b and c parameters were biased, the added difference over the unbiased condition was approximately equal to the sum of the differences attributable to bias in each parameter separately. In general the largest differences were with low discrimination tests. Tests with uniform b's showed little effect according to level of difficulty, but for peaked tests easy tests showed larger differences than did difficult tests. These results are show in Table 3.

In the case where both a and c were biased and the mean of the a bias quantity was zero, the differences in true score were approximately equal to the sum of the differences produced by the a and c parameters alone. This was also the case where the bias in a was attached to the Black group. When the bias was attached to Whites, however, both the a and c parameters for the item characteristics curves for Whites were changed and the differences appeared to interact. When the test was more discriminating for Whites, the observed differences were less than the sum of the separate effects for the two parameters; and when the test was more discriminating for Blacks, the differences were greater than the sum of the separate effects.

When both the a and b parameters were biased, there was generally an interaction such that again the effect was different from the sum of the separate effects. When the bias was attached to Blacks, there was a further interaction with the overall discrimination of the test. For tests with low discrimination, the added effect increased the difference beyond the separate

effects, particularly when the items were also more discriminating for Blacks. The largest difference added beyond the sum was .22 s.d. units for the easy test. For tests with high discrimination, however, the effect of bias in both parameters was to reduce the effect of the parameters separately. In highly discriminating, difficult tests where items were more discriminating for Whites, the combined effect was almost sufficient to wipe out the effect of the bias in b for both peaked and uniform distributions.

The addition of bias in the c parameter as well, so that bias was present in all three parameters, increased the magnitude of the differences but did not much change the patterns from the condition where just a and b were biased. Again, the effects were generally different from the sums of the separate effects. Notice, however, that some of the added differences were here becoming quite large. Particularly for relatively difficult, highly discriminating tests, the difference approximated an entire s.d. unit for both the peaked and uniform cases where bias was attached to the White group and items were more discriminating for Blacks. This is particularly interesting since the differences were smallest for this test under the b alone and the b and c conditions for the peaked tests and among the smallest for the uniform. The results are shown in Table 4.

Discussion

One of the purposes for this study was to determine if generalizable hypotheses could be generated concerning desirable characteristics of tests given that the actual amount and condition of bias in the tests is unknown. The results from this study are restricted in that only a limited set of test characteristics was investigated and generalizing to real testing situations

Table 4

Effects of Bias in Three Parameters on
Score Differences Between Blacks and Whites

Peaked Distribution

Mean Discrimination	Mean b	a for Whites			a for Blacks		
		same	a > W	a > B	same	a > W	a > B
Low	+.50	.69	.59		.80		.85
	0	.66	.48		.70		.85
	-.50	.66	.40		.66		.88
Medium	+.50	.65	.55	.82	.62	.55	.71
	0	.58	.43	.85	.54	.41	.67
	-.50	.55	.34	.89	.46	.30	.63
High	+.50	.59		.76	.58	.49	
	0	.60		.91	.42	.29	
	-.50	.64		1.05	.30	.12	

Uniform Distribution

Low	+.50	.69	.35		.74		.73
	0	.71	.17		.67		.80
	-.50	.76	.03		.63		.83
Medium	+.50	.54	.50	.63	.51	.45	.53
	0	.59	.51	.78	.48	.36	.54
	-.50	.66	.51	.94	.46	.28	.58
High	+.50	.43		.42	.46	.48	
	0	.51		.63	.37	.30	
	-.50	.65		.92	.29	.12	

[1] Bias quantity in first case is attached to both a & c parameters for Whites and on b parameter for Blacks; in second case on c parameter for Whites and for both a & b parameters for Blacks.

assumes that the IRT model used is in fact appropriate for a given test. Nonetheless, some usable hypotheses do seem to emerge which can now be tested in other studies in part by examining real test data. Since some of the characteristics of tests investigated here are often not specified in bias investigations, these may provide fruitful lines of study with existing bias data.

Some of the generalizations which may be worth further investigation are these:

1. In almost all instances the differences in true score between groups was larger with the peaked distribution than with the uniform distribution. The uniform distribution did not, of course, spread out the scores as well as the peaked distribution and other considerations would be appropriate in designing tests. In some instances, however, forming a more rectangular distribution may be a reasonable alternative.

2. Tests with relatively low discrimination on the average appear to produce larger differences than do tests with medium or high average discriminations.

3. Tests which are relatively easy for the high scoring group should have more rectangular distributions. Further investigation concerning the kinds of items which should be added with such tests is probably needed.

4. Tests which are quite difficult seem to produce smaller differences when they are peaked. Most of the very large differences due to bias occur with these types of tests, however, and their use with diverse groups should perhaps be questioned.

5. For tests with medium or high average discrimination, tests with difficulty centered between the higher and lower scoring groups appear

to be least likely to produce inflated differences due to bias. Only about 10 percent of the conditions showed high differences for these tests. This compared with 50 to 75 percent of the cases showing high differences with easy tests under the peaked b distributions. How to determine what this best level is in a given testing situation also needs further investigation.

The major implications from this study, however, concern the use of item bias techniques in improving the differences between two groups by removing biased items. It is known from previous research that some techniques are more sensitive than others to different types of bias in the items. This study has now demonstrated that not all types of bias are equally important in producing score changes. Fortunately, the most important single parameter for score change is item difficulty, for this is the aspect of biased items to which most of the techniques are reasonably sensitive. For those techniques which are insensitive to biases reflected in item discrimination or susceptibility to guessing, however, items which appear to be equally biased may not have the same importance with regard to score differences. Some errors where items fail to be detected may also be more serious than others. Moreover, the cases where bias lies in the discrimination are those where inflation of effect and some of the largest differences occur. The use of techniques which are insensitive to this source of bias should be weighed against the probability that bias of this nature is present in the test being examined.

My next steps in this research project will be to select five or six of the tests from the simulation study and generate item responses for the 500 hypothetical examinees in each group. I then plan to analyze these tests for bias using three or four bias techniques (depending on the costs) and find which items are successfully located by each of these techniques and which

are not. I then plan to look at different strategies for removing the biased items from the test and to evaluate the results by looking at the reduction in the score differences between groups and by comparing the original and adjusted scores with the unbiased true score on the test. We need to know these things if we are to successfully incorporate our item bias procedures into the test development process.

Another implication from these results is that the constant bias, those factors which contribute to bias across items in a consistent way and which were represented by the mean of the bias terms in this study, if these in fact exist, are vitally important to the score differences. Another of my own research directions at this time is to attempt to understand and define these relatively stable sources of bias in tests as opposed to the idiosyncratic sources often cited by critics or isolated in one of our item bias studies. I am just beginning a study where we will be attempting to verify certain hypotheses as to some of the sources of these kinds of biases.

I am optimistic that the research coming from the investigation of item bias is going to lead us in time to an understanding of test bias and some reasonable strategies for reducing its impact on test scores. Although our current state of understanding is still very primitive, I do not believe it is too soon to begin to try and form the link between the bias we are identifying in test items and the resultant impact on test scores.

References

Cole, N. S. Bias in Testing. American Psychologist, 1981, 36, 1067-1077.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

Pine, S.M., & Weiss, D.J. Effects of item characteristics on test fairness. Minneapolis: University of Minnesota, Psychometric Methods Program, Research Report 76-5, December 1976.

Rudner, L. M., Getson, P. R., & Knight, D. L. Biased item detection techniques. Journal of Educational Statistics, 1980, 5, 213-233.

Scheuneman, J. D. A new look at bias in aptitude tests. In P. Merrifield (Ed.), Abilities: Theories, Assessment, and Utility. San Francisco: Jossey Bass, in press.